ACS Code Replication

Henry Chen

July 7, 2023

Abstract

This report lists the difference and changes in replicated code as well as the corresponding results.

Contents

1 Introduction

In this task, I replicated the "15_ACS_Bartik" code. Specifically, I first proofread the code and checked whether there were any mistakes. In addition, I made some code optimizations on some functions as well as added the variable labels for interpretation. Furthermore, I tried new definitions, calculations, and methods to examine the results. A couple of extensions are also included. The report is structured into three parts: the first part highlights the fixes and changes made in the previous version, and the second part demonstrates the alternative methods I attempted. The last part is a summary of using H1B data for Bartik Instrument.

2 Proofread

2.1 Identified Errors

Here are the main errors that I identified in the code, listed in coding order:

1. Change "Place of Birth" value for the U.S.: The previous definition of place of birth variable "pob" for "United States" has a boarder range of definitions.

2. Change "Place of Birth" value Germany: The definition of place of birth variable "pob" for "Germany" is changed to incorporate more inclusive definitions.

3. Fix variable "num_inv_other": The previous variable also picks up values for those non-immigrant founded firms. Therefore, it inflates the variable itself. Meanwhile, it does not subtract the variable "num_inv_UnitedKingdom".

2.2 Fixes

For each fix mentioned above, I specifically deal with them as follows:

1. The variable "pob" for the United States is extended to include more U.S. territory areas. For instance, "Midway Islands" and "Wake Island". I modify the code to select a defined range (by IPUMS) of $[100, 12092]$ for U.S. territory.

2. The variable "pob" was originally using only "Germany" as the identifier for foreigners from Germany. I use the original detailed country code range $[45300, 45362]$ to capture the immigrants from Germany. The official website lists the range for Germany because it contains a unique code for "West Berlin" and "East Berlin" for the 1990 Census.

3. Based on the definition and calculation of the set of variables "num_inv_XXX", where XXX is a selected country or other, I understand these variables are defined as **the number of immigrant-founded firms who received any foreign investment from an assigned region, given within that city year**. Therefore, the variable "num_inv_other" should be also defined as **the number of immigrant-founded firms who received any foreign investment from the "other" region in that city year**. However, the original calculation of "num_inv_other" is

   $$\text{Total Count of Int'l Investment} - \sum^{N} \text{Count of Immig Firms with Int'l Investment from county j}$$

   where N is the set of countries selected, excluding "other" and "UK". In the correction, I fix two parts:

   - Restrict the calculation to only immigrant-founded firms
   - Subtract the count of firms with UK investment

   These two changes result in a *less* count of immigrant-founded firms with Int'l investment from other regions.

2.3 Modification in Variable: National Immigrant Population

There is one change in the variable calculation I made during the code. This, presumably, should not change the overall result dramatically, however, I point it out here in case we could trace back any change caused by this modification.

In the original code, line 277, we calculate the national-level immigrant population from each county by summing the city-level immigrant population, defined as "pop_from_XXXX". However, the variable "pop_from_XXXX" is the summing of the PUMA population of those PUMA intersecting with the given city. Thus, this calculation of the national immigrant population should be larger than the actual number.

I apply a different method to calculate the national-level immigrant population by directly performing the calculation on PUMA. Since we only need the national-level country-specific immigrant population each year, I take the original PUMA panel and sum the specific country immigrant population in that year to generate the national-level population. Then, I merge this national data onto the city level.

2.4 Code Optimization

I shall skip this part since the updated code is essentially generating the same results.

2.5 Comparison

Before diving into the new regression results, I post here the regression model as a reminder. The main regression model used in estimation is

$$\text{Has\_International\_Investor} = \alpha + \beta \times \log(\#\text{ of Immig. Founded Firms w/. Int'l Investment}) + \zeta + FE$$

where the instrument is $\log(\text{Bartik}_{t,2000})$, the Bartik instrument based on the Year 2000 immigrant share.

Here I list out a brief comparison between the original code and the updated one, as for examining the effect of whether the modification will incur significant variation. Table 1 shows the original results and Table 2 reports the modified results.

Table 1: Original Results

	OLS			2SLS		
x	0.0101*	0.0067	0.0075	0.1799**	0.1695*	0.0147
	(0.0053)	(0.0054)	(0.0054)	(0.0829)	(0.0921)	(0.0998)
Observations	39951	39481	39477	39815	39342	39346

Notes: Standard errors in parentheses.

Table 2: Modified Results

	OLS			2SLS		
x	0.0101*	0.0067	0.0075	0.1398**	0.1285*	0.0063
	(0.0053)	(0.0054)	(0.0054)	(0.0711)	(0.0780)	(0.0840)
Observations	39951	39481	39477	39940	39467	39471

Notes: Standard errors in parentheses.

While there is no difference in the OLS results comparing Table 1 and 2, the 2SLS shows some variations highlighted in red. Generally speaking, the magnitude of coefficients shrinks in the modified 2SLS results. And such dispersion should, I presume, comes from the correction of "num_inv_other", which downscales the instrument, and eventually reduces the size of the coefficient on instrumented variable x.

3 New Attempts

3.1 Calculating City-level Immigrant Population

The first attempt I did is the re-calculate the immigrant population at the city level. The previous methodology is summing up the PUMA-level immigrant population of those PUMA intersecting a given city. This could be biased toward larger PUMA or cities. If a city intersects with multiple PUMAs, then summing to get population will lead to a greater population in both this city and the nation. Potentially, it could result in an inflated share of immigrants attributed to larger cities and deflated one to sub-urban/rural/unincorporated areas.

To compare, I calculate both a simple and weighted average of the PUMA to proxy the city-level population. The weight is related to the count of a city's appearance in the crosswalk between the PUMA and the city, as well as the sum of the count of the appearance of all cities intersecting with this PUMA. For instance, if "New York City" appears 3 times in the PUMA-City crosswalk, that is to say, it intersects with three PUMA, and there are 2 more cities both with an appearance count at 1, then the share of population "New York City" takes from this PUMA is $\frac{3}{3+1+1}$.

I suppress the results of different city-level population calculations because the results are not significant and some suggest the opposite.

3.2 Weighted by City Features

Stemmed from Melanie's comment "*Endogeneity concern: native founders are located in cities where international investors would fund deals regardless of the presence of immigrants; e.g., large cities, Silicon Valley (partially accounted for by city fixed effects)*", should we add weights in the regression to control? If there is a magnet effect, then we should expect the effect to be larger but noisier if we look at the popular cities. I.e., look at a city with a higher count of native-founded firms or a total count of firms.

I report Table 3 for the regression analytically weighted by the count of firms in the city, and Table 4 is weighted by the city population.

Table 3: Attempt #1

	OLS			2SLS		
x	0.0074 (0.0131)	0.0025 (0.0133)	0.0007 (0.0134)	0.8618*** (0.3096)	0.9191*** (0.3408)	0.7912** (0.3313)
Observations	39951	39481	39477	39940	39467	39471

Notes: Standard errors in parentheses.

One phenomenon I am confused about regarding the results as well as the original ones is that the previous result from 2SLS with industry dummies changes so much compared to those without industry dummies. The OLS, on the other hand, did not change too much. In comparison, the results in Table 3 and 4 do not exhibit significant changes.

Table 4: Attempt #2

	OLS			2SLS		
x	0.0130 (0.0079)	0.0118 (0.0081)	0.0113 (0.0081)	0.6411** (0.2536)	0.6254*** (0.2377)	0.5185** (0.2267)
Observations	39951	39481	39477	39940	39467	39471

Notes: Standard errors in parentheses.

3.3 Split Sample by City's Firm Count

In this section, I test whether there are differential effects between entrepreneurship-intensive and less entrepreneurship-intensive areas. An entrepreneurship-intensive city is defined as having the count of firms in the city greater than the median of the distribution of firm count in the city. There, as expected, would be fewer entrepreneurship-intensive cities than those no-entrepreneurship-intensive ones (a brief count: 31 cities are classified as entrepreneurship-intensive, while 3040 are the opposite).

I have two hypotheses on the so-called "city size" effect on the probability of a native-founded firm receiving an international investment. The first anticipation is that, for a larger city, it is more likely that a firm (native-founded) to receive international funding because either international investors will have better access to larger cities, more network effect, stronger magnet from immigrant-founded firms to those native-founded, or the investor applies more "spread and pray" strategy. Or, from a sticky investment choice perspective, if a city has more entrepreneur firms, then regardless it has immigrants or not, it is more likely to get funded and so is an international investment.

The counter hypothesis to the above is that, if for instance, the funding is limited or firms crowd into big cities for tax purposes or else, the probability for a firm to receive international funds will be lower from the statistical smaller. Or to say, lemons are more possible and a good idea is hard to find, then as a result, the posterior distribution could draw a smaller possibility of getting international funding.

Nevertheless, I am more toward the first hypothesis where big cities should have a larger effect. Based on that, I rerun the regression conditional on whether the city is classified as a large city or not. Table 5 and 6 report the results for the large-city sample and the small-city sample. The OLS results from both tables do not suggest that there are differential effects between small and large cities. However, the 2SLS regressions show that, for large cities, native-founded firms enjoy a magnet effect from immigrant-founded firms on receiving international funding, compared to companies in smaller cities.

Table 5: Attempt #3: Large Cities

	OLS			2SLS		
x	0.0084 (0.0092)	0.0038 (0.0095)	0.0053 (0.0095)	0.5843** (0.2411)	0.6243** (0.2767)	0.5822** (0.2761)
Observations	28916	28561	28561	28916	28561	28561

Notes: Standard errors in parentheses.

3.4 Changing the Instrumented Variable

One thing I notice in the panel data constructed is that the original instrumented variable $x \equiv$ log_tot_winv_immig_city has around one-third of the observations with missing values, while the instrument is only 0.55% missing. I switch the instrumented variable from log(tot_winv_immig_city) to log(1 + tot_winv_immig_city), and the original regression results, replicating Table 1, is shown in Table 7

Table 6: Attempt #4: Small Cities

	OLS			2SLS		
x	0.0058 (0.0086)	0.0055 (0.0088)	0.0082 (0.0088)	0.0200 (0.1144)	-0.0312 (0.1142)	0.7733 (0.5002)
Observations	11035	10817	10813	11024	10803	10807

Notes: Standard errors in parentheses.

Table 7: Changed Instrumented Variable

	OLS			2SLS		
x	0.0157*** (0.0041)	0.0158*** (0.0041)	0.0143*** (0.0042)	0.0797*** (0.0295)	0.0821*** (0.0314)	0.0823*** (0.0314)
Observations	55259	54670	54570	54995	54314	54410

Notes: Standard errors in parentheses.

As a comparison, I run the regression on samples where the original instrumented variable log(tot_winv_immig_city) is missing in value. Table 8 reveals a similar coefficient estimation as in Table 1

Table 8: Changed Instrumented Variable on Original Sample

	OLS			2SLS		
x	0.0148** (0.0061)	0.0109* (0.0063)	0.0100 (0.0063)	0.0888** (0.0449)	0.0789* (0.0477)	0.0040 (0.0530)
Observations	39951	39477	39481	39940	39467	39471

Notes: Standard errors in parentheses.

This is worth bringing up if we consider the missing values causing a false estimation.

3.5 Extension

In this part, I append more required results to formalize our understanding of the effective channels. Upon the previous regression results, I will additional show three aspects:

1. First Stage Regression Results for L/S city regressions
2. Lagged Instrument and Instrumented Variables
3. Replace Instrumented Variable as Share of Immigrant-Founded Firms under certain Features

$$\text{Instrumented} \equiv \frac{\text{\# Immigrant-founded Firms w/. Feature X}}{\text{(\# Immigrant-founded Firms w/. Feature X + \# Native-founded Firms w/. Feature X)}}$$

3.5.1 First Stage Regression Results

In this short section, I report the regression results for the first stage in IV regressions. These IV regressions come from the preliminary regression on classifying samples into **Large Cities** or **Small Cities**. Following the same order as previous, Table 9 exhibits the first stage results for Large and Small cities.

Table 9: First Stage Regression

	Large			Small		
z	0.1619*** (0.0232)	0.1443*** (0.0233)	0.1430*** (0.0233)	-0.6005*** (0.0763)	-0.6192*** (0.0788)	-0.0913** (0.0390)
Control	X	X	✓	X	X	✓
Year, City	✓			✓		
YearXFound Year, City		✓			✓	
YearXFound Year, HQ			✓			✓

Notes: Standard errors in parentheses.
Notes: Control for Industry Dummies.

The results for large and small cities suggest the opposite conclusions. I think it mainly comes from the fact that entrepreneurship tends to happen in large cities whereas the small cities also enjoy the immigrant inflow, but just not for the purpose of founding a new company.

3.5.2 Lagged Instrument & Instrumented

In this part, I test the one-year lagged and the two-year lagged variables. In detail, I test the following combinations:

1. both one-year lagged instrumented and instrument
2. both two-year lagged instrumented and instrument
3. one-year lagged instrumented and two-year lagged instrument
4. current-year instrumented and one-year lagged instrument

Besides these combinations, I also test the regression conditional on large cities only. All regression results do not show any significance. I suppress the regression tables here for briefness.

3.5.3 Replacing with Share of Immigrant-Founded Firms w/. Features

I construct the Share variable as

$$\text{Instrumented} \equiv \frac{\text{\# Immigrant-founded Firms w/. Feature X}}{(\text{\# Immigrant-founded Firms w/. Feature X} + \text{\# Native-founded Firms w/. Feature X})}$$

where the Feature is conditional on receiving an international investment. Using this variable for replacing the previous instrumented variable, I present Table 10 for the estimation results.

Table 10: New Instrumented Variable

	OLS			2SLS		
x	-0.0937*** (0.0071)	-0.0941*** (0.0072)	-0.0955*** (0.0072)	0.6796* (0.3901)	0.6582 (0.4481)	0.0561 (0.7515)
Observations	39951	39481	39477	39940	39467	39471

Notes: Standard errors in parentheses.

The OLS results suggest the opposite effect compared with the 2SLS regression. One explanation that I could come up with is that the OLS is still more of a native regression. It is basically only a reflection of the endogenous variable, which more share of immigrant-founded firms getting international investment is equivalently saying less share of native-founded firms getting international investment.

The first result in 2SLS, on the other hand, suggests a similar idea as the previous regression. The more immigrants flooding into the city, the more possible it brings more immigrant-founded firms (together with foreign investment), and the more magnet effect onto those native-founded firms. (Their first stages are significantly positive)

3.6 New Extension

Based on the comments from Melanie, I append four additional regression results here for records.

1. Try share = # immig firms w/ international capital / (# immig firms + # natives firms); i.e., denominator unconditional on raising international capital
2. Try share = # immig firms/ (# immig firms + # native firms); i.e., numerator and denominator unconditional on raising international capital
3. Try current Table 10, lagging instrumented and instrument (this will get rid of mechanical issues)
4. Try current Table 10, controlling for log(# immig firms + # native firms) to get a measure of the current size of entrepreneurship

Table 11, 12, 13, and 14 report the estimation results for each extension above.

Table 11: Share #2 Regression

	OLS			2SLS			2SLS: FS		
x	0.0178***	0.0153***	0.0162***	0.2154*	0.1922	0.0409			
	(0.0058)	(0.0058)	(0.0059)	(0.1104)	(0.1172)	(0.5459)			
z							0.2569***	0.2444***	0.0429**
							(0.0242)	(0.0244)	(0.0204)
Obs	39951	39481	39477	39940	39467	39471	39940	39467	39471

Notes: Standard errors in parentheses.

Table 12: Share #3 Regression

	OLS			2SLS			2SLS: FS		
x	0.0127*	0.0101	0.0126*	0.2758**	0.2833*	-0.6616			
	(0.0069)	(0.0070)	(0.0070)	(0.1356)	(0.1452)	(0.4235)			
z							0.1895***	0.1794***	-0.0344***
							(0.0170)	(0.0171)	(0.0090)
Obs	46453	45902	45897	46348	45799	45804	46348	45799	45804

Notes: Standard errors in parentheses.

4 Using LCA Data

In this part, I replicate the methods from Kerr and Lincoln (2010) to use Labor Condition Application (LCA) data for constructing Bartik Instrument and further compare the results with those from ACS. The H1B data, on the other hand, is postponed due to the pending request for access to USCIS. I will perform a similar analysis using H1B data as soon as USCIS responds.

The benefit of using LCA to proxy the pool of H1B is more-or-less reasonable in the sense that each employer is required by law to submit an LCA before the worker applies for an H1B visa. It creates an approximation of how many immigrants a firm hire in a fiscal year. However, the literature also noted that the LCA is not strictly a one-to-one relationship with the H1B.

Table 13: Lagged Share #1 Regression

	OLS			2SLS			2SLS: FS		
lx	0.0095	0.0090	0.0100	0.3984	0.4112	0.3342			
	(0.0073)	(0.0074)	(0.0074)	(0.2813)	(0.3411)	(0.3601)			
lz							0.0993***	0.0832***	0.0769***
							(0.0200)	(0.0202)	(0.0201)
Obs	33717	33391	33390	33709	33382	33383	33709	33382	33383

Notes: Standard errors in parentheses.

Table 14: Share #1 Regression, Control Current Size

	OLS			2SLS			2SLS: FS		
x	-0.0955***	-0.0963***	-0.0977***	0.5915*	0.5716	0.1072			
	(0.0071)	(0.0072)	(0.0072)	(0.3163)	(0.3609)	(0.3908)			
z							0.0980***	0.0865***	0.0606***
							(0.0196)	(0.0198)	(0.0165)
Obs	39951	39481	39477	39940	39467	39471	39940	39467	39471

Notes: Standard errors in parentheses.

To clarify the use of LCA, I highlight some features of the LCA that would affect the method to construct the data panel.

First, an LCA is usually associated with the H1B application. The employer submits an LCA that specifies the length of time for employment and an H1B petition for a worker if the employer wants to hire that worker. When an H1B is granted, the H1B is generally valid for the same length of time as the LCA. If the employer wants to extend employment beyond the length of time filling in the LCA, then the employer needs to resubmit a new LCA and a new H1B petition. In a nutshell, an LCA is generally linked to an H1B.

Furthermore, when an H1B petition is not selected, the associated LCA goes expired and cannot be reused for a future application. In other words, a new LCA is needed for a worker when continuing to apply for H1B in the following fiscal year.

Other than H1B, the LCA could also be used for applying for E-3 Visa and H-1B1 Visa. Generally speaking, an LCA is used by the Department of Labor (DOL) to ensure that firms hiring foreign workers will not aggravate the working conditions and wages of U.S. workers. When there are more applications for E-3 and H-1B1 Visa, the approximation for H1B could be mitigated.

Lastly, the performance of using LCA for proxying the H1B could be influenced by the total H1B applications and H1B cap in a given year. Since LCA is a prior H1B petition count, when the H1B applications exceed the H1B cap, LCA cannot reflect the distribution of those approved H1B applications.

Nevertheless, LCA could provide some insights into whether ACS results persist across different data sources. This section is structured as follows: Part one gives an introduction to the data used for analysis. Part two presents a brief summary of statistics for the immigrant population and the LCA count. Part three exhibits the regression results using the LCA-Bartik instrument. Part four concludes.

4.1 Data

Using the same data as Kerr and Lincoln (2010), I download the LCA from DOL for 2008 to 2022. All variables and regression setting are the same as before, except for the Bartik variable. Since the H1B data does not specify the applicant's nationality, I modify the Bartik variable calculation as follows: given the set of countries $\{C\}$ used in previous calculation (i.e., India, UK, Canada, Israel, France, Australia, China, Germany, Russia, Italy, Spain, and Other), the Bartik variable at time t in city c is

$$\text{Bartik}_{t,c} = \frac{\sum_i^{\{C\}} pop_{i,c}}{\sum_i^{\{C\}} pop_{i,n}} \times \mathbf{g}_{H1B,t,c}$$

where $\sum_i^{\{C\}} pop_{i,n}$ is the number of defined immigrant population at the national level, and $\mathbf{g}_{H1B,t,c}$ is the H1B annual population growth rate in city c

4.2 Preliminary Results

The regression results are shown in Table 15. The first three columns are OLS results, the second threes are second stage from 2SLS, and the last three are the first stage from 2SLS.

Table 15: H1B Regression

	OLS			2SLS			2SLS: FS		
x	0.0064	0.0040	0.0049	-0.1628*	-0.1667*	-0.1740*			
	(0.0065)	(0.0066)	(0.0066)	(0.0979)	(0.0997)	(0.0991)			
z							-0.0352***	-0.0349***	-0.0350***
							(0.0023)	(0.0024)	(0.0024)
Obs	36283	35877	35874	24274	24036	24036	24274	24036	24036

Notes: Standard errors in parentheses.

References

Kerr, W. R., & Lincoln, W. F. (2010). The supply side of innovation: H-1b visa reforms and us ethnic invention. *Journal of Labor Economics*, *28*(3), 473–508.